

# Essais contrôlés randomisés (ECR)

Howard White, Shagun Sabarwal et Thomas de  
Hoop

## CENTRE DE RECHERCHE DE L'UNICEF

Le Centre de recherche Innocenti est le bureau de l'UNICEF spécialisé en recherche scientifique. Les objectifs prioritaires du Centre de recherche sont d'améliorer la compréhension internationale des questions liées aux droits des enfants afin de faciliter la pleine application de la Convention relative à ces droits dans le monde entier. Le Centre a pour objectif de mettre en place un cadre intégré pour la gestion de la recherche et des connaissances au sein de l'organisation, afin de soutenir ses programmes et politiques à l'échelle mondiale. Les publications du Centre, en présentant un vaste éventail d'opinions, contribuent au débat international sur les questions liées aux droits des enfants.

Les opinions exprimées n'engagent que leurs auteurs et/ou réviseurs et sont publiées afin d'encourager le dialogue sur les méthodes d'évaluation d'impact. Cette publication ne reflète pas nécessairement les politiques ou points de vue de l'UNICEF. **NOTES MÉTHODOLOGIQUES DU CENTRE DE RECHERCHE**

Les notes méthodologiques du Centre de recherche de l'UNICEF ont le but de partager des pratiques, des méthodes et des modèles ainsi que des recommandations de chercheurs et analystes reconnus. Elles sont dirigées principalement au personnel de l'UNICEF qui conduit des recherches ou qui interprète les résultats et analyses des évaluations réalisées par des chercheurs externes pour les prises de décision en phase de programmation ou à l'appui des politiques ou de la sensibilisation.

Cette note méthodologique a suivi un procès interne de révision par pairs.

Ce texte n'a pas été édité conformément aux standards officiels de publication et UNICEF décline toute responsabilité pour les possibles erreurs.

La reproduction de parties de cette publication est permise à condition que la source soit dûment citée. Pour l'utilisation d'une partie substantielle ou de la totalité de la publication, veuillez bien vous adresser au Département de Communication au suivant courriel [florence@unicef.org](mailto:florence@unicef.org)

Pour consulter ou télécharger ces notes méthodologiques, visitez <http://www.unicef-irc.org/KM/IE/>

Nous conseillons d'utiliser la citation suivante pour toute référence au document présent :

White, H., Sabarwal S. et de Hoop, T. (2014). Essais contrôlés randomisés (ECR), *Note méthodologique n° 7*, Centre de recherche Innocenti, Florence.

**Remerciements** : Cette note méthodologique a profité des conseils de beaucoup d'individus. L'auteur et le Centre de recherche désirent remercier tous ceux qui ont contribué et en particulier :

**Pour leur contribution** : Dugan Fraser, Simon Hearn, Patricia Rogers, Jessica Sinclair Taylor

**Pour leur révision** : Nikola Balvin, Samuel Bickel, Sarah Hague, Sudhanshu Handa, Aisha Yousafzai

© Fonds des Nations Unies pour l'enfance (UNICEF), septembre 2014

Centre de recherche Innocenti de l'UNICEF

Piazza SS. Annunziata, 12

50122 Florencia (Italia)

Tél. : (+39) 055 20 330

Fax : (+39) 055 2033 220

[florence@unicef.org](mailto:florence@unicef.org)

[www.unicef-irc.org](http://www.unicef-irc.org)

## 1. ESSAIS CONTRÔLÉS RANDOMISÉS : UNE BRÈVE DESCRIPTION

L'essai contrôlé randomisé (ECR) est une méthode suivie dans le cadre de l'évaluation d'impact. À partir d'une population admissible, cette méthode permet de sélectionner de façon aléatoire le groupe expérimental qui bénéficiera d'une intervention programmatique ou politique et le groupe contrôle qui servira de point de comparaison. L'ECR permet d'évaluer dans quelle mesure les impacts escomptés sont obtenus.

Lors d'un ECR, le programme ou la politique est considéré(e) comme une « intervention » qui consiste à évaluer au moyen d'un ensemble d'indicateurs prédéfini la capacité d'un « traitement » (à savoir, les éléments du programme ou de la politique à l'étude) à remplir ses objectifs. L'ECR présente l'avantage d'établir solidement les liens de causalité, ce qui permet aux évaluateurs et aux agents d'exécution du programme de confirmer que c'est bien de l'intervention que découlent les résultats obtenus.

L'ECR est une forme [expérimentale](#) de l'évaluation d'impact. Les formes [quasi expérimentale](#) et [non expérimentale](#) de l'évaluation d'impact sont abordées dans les autres notes méthodologiques de cette série : la Note n° 6, Présentation des stratégies d'attribution causale, la Note n° 8, Méthodes et modèles quasi expérimentaux et la Note n° 9, Études de cas comparatives.

L'ECR se caractérise principalement par le fait que les membres de la population admissible à un traitement sont répartis de façon aléatoire dans un ou plusieurs [groupes expérimentaux](#) (qui reçoivent le traitement de l'intervention<sup>1</sup> ou ses variantes) ou dans un groupe contrôle (qui ne fait l'objet d'aucune intervention ou, si le traitement est une version améliorée d'une intervention existante, continue de bénéficier de l'intervention habituelle). Après un délai préalablement fixé, les effets observés chez les différents groupes et dans plusieurs domaines d'impact font l'objet d'une comparaison. L'encadré 1 explique la différence entre la randomisation et l'échantillonnage aléatoire, deux caractéristiques clés de l'ECR.

### Encadré 1. La randomisation et l'échantillonnage aléatoire

La randomisation et l'échantillonnage aléatoire sont deux techniques distinctes. L'échantillonnage aléatoire désigne la façon dont un échantillon est composé à partir d'une ou plusieurs populations, tandis que la randomisation désigne la façon dont les individus ou les groupes sont inclus dans un groupe expérimental ou dans un groupe contrôle. Généralement, les ECR utilisent à la fois l'échantillonnage aléatoire (pour pouvoir faire des déductions sur une population) et la randomisation (qui caractérise l'ECR).

Le modèle d'ECR le plus simple inclut un groupe expérimental (ou bras expérimental) et un groupe contrôle. Des variations sont aussi possibles en présence des éléments suivants :

- une situation avec **plusieurs bras expérimentaux** : par exemple un premier groupe expérimental bénéficie de l'intervention A et un deuxième groupe expérimental bénéficie de l'intervention B, ou
- **un plan factoriel**, dans lequel un troisième bras expérimental bénéficie à la fois des interventions A et B.

<sup>1</sup> L'ECR peut servir à mesurer les interventions programmatiques (p. ex., distribution de suppléments nutritionnels dans le cadre d'un programme de nutrition) ou politiques (p. ex., distribution d'argent dans le cadre d'une politique favorable aux transferts de fonds). Par souci de concision, les termes « programme » ou « intervention » font référence à ces deux types d'intervention dans cette note.

Lorsqu'une intervention est déjà en place, il est préférable que le groupe contrôle continue d'en bénéficier. L'ECR montrera alors les différences entre la nouvelle intervention et l'intervention originale.

Dans le cas d'un ECR simple, la même unité d'analyse est utilisée pour l'intervention et la randomisation. Par exemple, lors de l'évaluation d'un programme nutritionnel, les individus recevant des suppléments nutritionnels pourront être sélectionnés au moyen d'un processus de randomisation.

Pour des raisons pratiques et éthiques, il est néanmoins plus courant d'utiliser un [modèle d'ECR en grappes](#), dans lequel chaque unité d'étude est composée de plusieurs unités bénéficiant de l'intervention. Par exemple, les interventions éducatives sont généralement mises en œuvre à l'échelle d'un établissement dans son ensemble, même si l'intervention est en réalité axée sur l'enseignant, la classe ou chaque élève et que les effets sont mesurés à l'échelle de ce dernier. De même, les interventions nutritionnelles peuvent s'appliquer à l'échelle de la communauté ou du sous-district. L'ECR en grappes est généralement mieux adapté dans le contexte de l'UNICEF, ses programmes étant souvent de grande envergure.

### Points principaux

1. L'ECR mesure les conséquences d'une intervention programmatique ou politique sur un effet direct donné.
2. L'ECR se caractérise par le recours à la **randomisation** des bénéficiaires de l'intervention. On le qualifie de modèle expérimental (ou méthode expérimentale).
3. L'ECR ne permet de mesurer l'impact que dans certains cas. Il faut notamment que l'échantillon d'étude soit important, que les impacts attendus de l'intervention programmatique ou politique soient faciles à définir et à mesurer (p. ex., réduction des retards de croissance), et que l'ECR soit planifié avant le début de l'intervention.

## 2. QUAND UTILISER LES ESSAIS CONTRÔLÉS RANDOMISÉS ?

### L'ECR doit être planifié dès le début du programme

L'ECR doit être planifié dès le début de la mise en œuvre du programme, et la participation au programme doit être soigneusement contrôlée, sans jamais perdre de vue l'essai en cours. L'ECR ne peut être mis en place rétrospectivement.

Seule la [méthode d'encouragement](#) échappe à cette règle. Dans ce cas, la randomisation des participants ne porte pas sur une intervention, mais sur l'accès à des documents promotionnels ou à un supplément d'information sur les avantages de l'intervention à leur disposition, en vue de renforcer la participation à cette intervention. Les méthodes d'encouragement peuvent être utilisées lorsque des programmes sont accessibles à tous, mais qu'ils ne sont pas adoptés par tous.

### L'ECR requiert un échantillon important

L'ECR peut être mené uniquement si l'échantillon suffit à identifier précisément les effets du programme. Le modèle d'étude doit présenter une « puissance statistique » suffisante.

Ce terme propre à la statistique désigne la probabilité de conclure, sans erreur possible, que le programme en place fonctionne. L'élaboration d'un ECR consiste notamment à calculer cette puissance statistique de

façon à trouver la taille d'échantillon qui suffira à estimer l'impact du programme (voir l'encadré 2). Plus l'échantillon est élevé, plus la puissance statistique est importante.

Dans le cas de l'ECR en grappes, la puissance statistique de l'étude repose sur le nombre de grappes plutôt que sur le nombre d'observations. Par exemple, un échantillon composé de 50 communautés dans lesquelles cinq ménages sont étudiés (soit 50 grappes) présente une puissance statistique nettement plus élevée qu'un échantillon composé de 25 communautés dans lesquelles 10 ménages sont étudiés (soit 25 grappes) — et cela même si l'échantillon porte sur 250 ménages dans les deux cas.

Des outils logiciels permettent d'effectuer ces calculs de puissance statistique (p. ex., Optimal Design), mais il est préférable de confier cette tâche à une personne expérimentée en la matière.

### Encadré 2. Calculs de puissance statistique

La puissance statistique désigne la probabilité de discerner un impact lorsque le programme en a un. Pour calculer la puissance statistique et la taille d'échantillon requise pour réaliser une évaluation, les évaluateurs formulent généralement des hypothèses sur la taille d'effet escomptée, le niveau de pertinence statistique et la corrélation intragrappe (concernant l'ECR en grappes). La corrélation intragrappe est une méthode de statistique descriptive indiquant, sur une échelle de 0 à 1, dans quelle mesure les groupes (p. ex., les ménages) ou les individus au sein d'une même grappe ont un comportement similaire. Plus la corrélation intragrappe est élevée, plus l'échantillon doit être important. Dans le cas de l'ECR en grappes, l'augmentation de la puissance statistique dépend généralement plus de la multiplication des grappes que de celle des individus ou des groupes inclus dans les grappes.

### L'ECR doit faire suite à une étude ou évaluation formative

Dans la plupart des cas, il est souhaitable d'attendre que la conception du programme soit terminée et validée, voire que le programme ait atteint sa maturité, pour mettre en place un ECR. Cette règle s'applique aussi aux essais cliniques de médicaments, ceux-ci étant menés après une phase de mise au point et de tests initiaux. L'étude formative ou l'analyse de situation doit permettre d'évaluer les facteurs sous-jacents du problème abordé par le programme (p. ex., résultats scolaires faibles) pour ensuite orienter sa conception. Par exemple, les mesures visant à renforcer la fréquentation scolaire des élèves sont inutiles si l'enseignant est lui-même souvent absent.

L'évaluation formative, qui se présente sous la forme d'une étude pilote ou d'une démonstration de faisabilité par exemple, évalue s'il est possible de mettre en œuvre le programme, tout en garantissant son utilisation par les bénéficiaires, et améliore la qualité de la mise en œuvre. Les résultats de l'étude pilote peuvent permettre d'identifier les modifications devant être apportées lors de la conception de la version élargie qui suivra l'étude pilote. Même si l'ECR est parfois utilisé comme démonstration de faisabilité, il est déconseillé d'y avoir recours à un stade aussi précoce. Afin d'éviter de gaspiller des ressources limitées et de formuler des constatations erronées, il convient d'éviter d'avoir recours à un ECR — très coûteux — pour évaluer un programme en cours de conception.

### L'ECR doit être adapté à la nature du programme évalué

L'ECR convient surtout aux programmes qui visent à produire des impacts clairs et mesurables pouvant être attribués à une intervention particulière (ou un ensemble d'interventions particulières) et qui peuvent

faire l'objet d'une [analyse de dépendance](#). L'ECR n'est pas adapté aux programmes émergents ou visant des résultats difficiles à mesurer.

Une étude a identifié quatre situations dans le cadre desquelles la randomisation n'est pas souhaitable ou réalisable<sup>2</sup> : (1) lorsqu'il est nécessaire d'obtenir des réponses rapidement ; (2) lorsque le besoin de précision et la question de causalité ne sont pas prioritaires ; (3) lorsqu'il est impossible de maîtriser la randomisation (p. ex., lorsque la question de causalité à l'étude nécessite de s'exposer à une situation indésirable) ; et (4) lorsque le travail empirique préliminaire qui a été accompli n'est pas suffisant et que l'intervention ou le programme n'a pas encore atteint sa maturité.

Les programmes qui ne se prêtent pas à la randomisation incluent notamment ceux dont l'intervention touche un faible nombre d'unités expérimentales (p. ex., le soutien apporté par une institution à une organisation unique) et ceux dont les activités et les effets directs attendus n'ont pas été définis à l'avance. Avec un peu d'imagination, il est toutefois parfois possible d'adapter un programme qui, a priori, ne se prêtait pas à la randomisation. Par exemple, la méthode d'encouragement peut permettre l'évaluation d'un programme national. De même, dans le cadre d'un programme de réforme institutionnelle, les politiques salariales peuvent faire l'objet d'un ECR portant sur les différentes mesures d'encouragement proposées aux travailleurs.

Lorsque l'ECR ne peut être réalisé, les décideurs concernés doivent en être informés. Dans ce cas, la réalisation de l'évaluation d'impact peut passer par une [méthode quasi expérimentale](#) (telle que l'[appariement sur le score de propension](#) ou ASP) ou par une **Error! Hyperlink reference not valid.** (telle que la [reconstitution de processus](#)) (voir les Notes n° 8, Méthodes et modèles quasi expérimentaux et n° 6, Présentation des stratégies d'attribution causale.)

Lorsque l'évaluation d'impact ne peut pas encore être menée, il peut être possible et utile de mener une évaluation à un niveau inférieur de la chaîne de causalité, par exemple sur les produits et les effets directs. L'évaluation des effets directs permet de recueillir des informations sur les résultats à court terme qui sont étroitement liés aux impacts attendus. Si la chaîne de causalité est bien définie et bien comprise, ce type d'évaluation peut être pertinent.

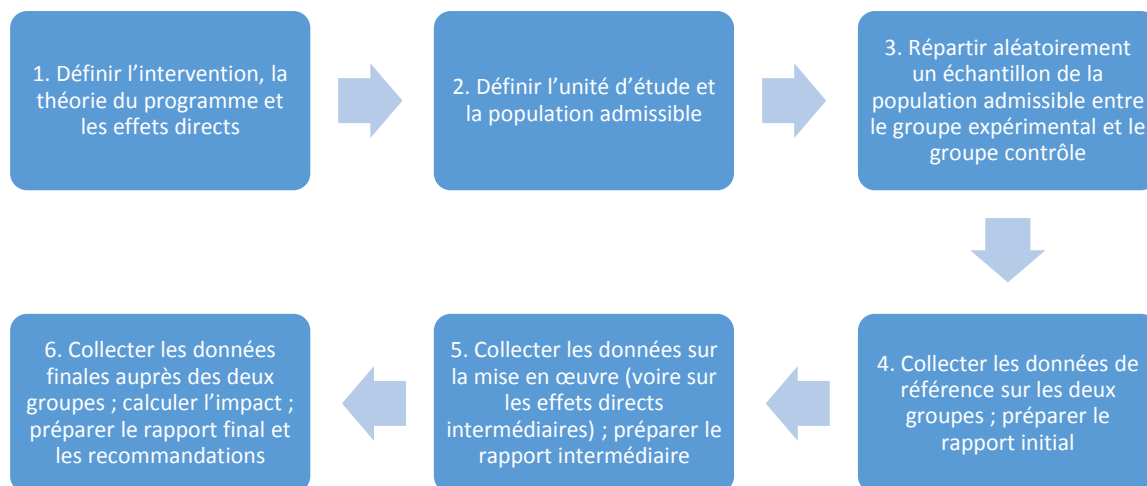
Lors d'une campagne de vaccination, par exemple, il peut être plus utile de rendre compte du statut vaccinal au moyen d'une évaluation que de collecter ultérieurement des données sur la morbidité ou la mortalité. Une évaluation de processus peut aussi être réalisée afin d'améliorer la mise en œuvre et d'accroître l'utilité de l'évaluation d'impact ultérieure.

---

<sup>2</sup> Shadish, W. R. *et al.*, 2002, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston

### 3. COMMENT MENER UN ESSAI CONTRÔLÉ RANDOMISÉ ?

Figure 1. Déroulement d'un essai contrôlé randomisé



#### 1. Définir l'intervention, la théorie du programme et les effets directs

Comme pour toute évaluation d'impact, l'ECR doit commencer par définir clairement l'objet et les raisons de l'évaluation, de même que les effets directs et les impacts attendus. À cette fin, il est très utile d'élaborer une [théorie du changement](#) qui expose clairement les modifications que le programme entend apporter et les activités qui seront mises en œuvre dans ce but (voir la Note n° 2, Théorie du changement). L'analyse de la théorie du changement permet d'identifier les questions d'évaluation relatives à la chaîne de causalité ainsi que les impacts à évaluer. À ce stade, il est également utile de préciser comment seront utilisées les données collectées dans le cadre de l'ECR.

#### 2. Définir l'unité d'étude et la population admissible

Comme indiqué plus haut, la population admissible et l'unité d'étude visées par la randomisation doivent être clairement identifiées lors de l'élaboration de l'ECR, afin d'éviter toute incohérence. L'unité d'étude doit faire l'objet d'une attention particulière : la randomisation (c'est-à-dire, l'affectation aléatoire dans le groupe expérimental ou le groupe contrôle) doit-elle porter sur des individus ou sur des groupes d'individus (p. ex., des écoles ou des villages) ? Au début de l'étude, les évaluateurs doivent aussi identifier les sous-groupes présentant un intérêt potentiel de façon à s'assurer que l'étude présentera une « puissance statistique » suffisante pour permettre la réalisation des analyses d'intérêt sur ces sous-groupes.

### 3. Répartir aléatoirement un échantillon de la population admissible entre le groupe expérimental et le groupe contrôle

Les ECR peuvent être préparés de plusieurs façons. Différentes méthodes existent pour mener un ECR sur un programme, et divers modèles d'ECR peuvent aussi être utilisés en fonction des caractéristiques du programme. Les trois principaux modèles sont décrits ci-dessous.

**La randomisation échelonnée** signifie que le programme couvrira progressivement toutes les unités d'étude s'il est jugé efficace. Dans ce cas, la méthode s'applique à la date d'intégration au programme. Comme les organismes d'exécution déploient souvent le programme en plusieurs étapes, il est possible de sélectionner aléatoirement l'ordre selon lequel les participants en bénéficieront. Par exemple, lorsque des contraintes budgétaires et logistiques empêchent de déployer immédiatement le programme dans tout le pays, les unités qui bénéficieront du programme lors de la première étape peuvent être sélectionnées aléatoirement. Cette approche a notamment été adoptée par le programme mexicain de transfert conditionnel de fonds, intitulé « Progres/Oportunidades »<sup>3</sup>. Lors de la phase initiale, une version pilote du programme a été menée auprès de 506 communautés, dont la moitié a bénéficié immédiatement du programme et l'autre moitié a joué le rôle de groupe contrôle pendant deux ans. Autrement dit, les communautés ont été aléatoirement réparties en deux groupes. Le premier groupe a bénéficié du programme dès la première année, tandis que le second groupe a commencé à en bénéficier lors de la troisième année (après avoir servi de groupe contrôle pendant les deux premières années).

**La randomisation avec rehaussement du seuil** est utilisée lorsque les ressources budgétaires disponibles ne suffisent pas à couvrir l'ensemble de la population admissible. Comme le programme ne couvrira pas toutes les unités de la population admissible, le recours à la randomisation est souvent la méthode la plus équitable et la plus transparente pour sélectionner les bénéficiaires. Une telle approche est généralement adoptée lorsque l'admissibilité est limitée par un seuil, par exemple le seuil de pauvreté. Si la population admissible ne peut être entièrement couverte par les budgets disponibles, il est possible de rehausser légèrement le seuil pour permettre la randomisation. Par exemple, si seuls les ménages ayant des enfants âgés de 24 mois peuvent participer à un programme de nutrition, le seuil peut être rehaussé à 30 mois. La même approche peut être appliquée à l'échelle géographique. Un programme prévu pour 50 communautés peut tout d'abord identifier 100 communautés, puis en sélectionner aléatoirement 50 qui participeront au programme. Dans ce dernier cas, le recours à la randomisation par appariement (c'est-à-dire la constitution de paires au sein des communautés et l'attribution aléatoire d'un élément de chaque paire au groupe expérimental) renforcerait la méthode employée.

**Les méthodes d'encouragement** sont utilisées lorsque les programmes et les politiques sont accessibles à tous, mais qu'ils ne sont pas adoptés par tous. Le groupe expérimental est encouragé à avoir recours à l'intervention. Toutefois, ces mesures d'encouragement et leurs effets ne doivent pas influencer l'intervention. Par exemple, la méthode d'encouragement peut consister à réaliser des campagnes d'information sur un programme en cours dans certains villages, mais pas dans d'autres. Les villages où les campagnes seront réalisées sont sélectionnés de façon aléatoire parmi tous les villages où le programme a été mis en œuvre. Les chercheurs mesurent ensuite l'impact du programme par rapport aux effets directs escomptés en comparant les effets directs sur les villages contrôles et sur les villages expérimentaux (ici, les villages ayant bénéficié de la campagne d'information). Cette approche permet d'évaluer l'impact en comparant le taux d'utilisation du programme dans les villages couverts par la campagne d'information et les autres.

Plusieurs méthodes existent pour répartir aléatoirement les groupes de population entre le groupe expérimental et le groupe contrôle :

- **Randomisation simple** : les individus ou les sites sont répertoriés puis répartis entre le groupe expérimental et le groupe contrôle au moyen d'un générateur de nombres aléatoires.

---

<sup>3</sup> Ministère mexicain du développement social (SEDESOL), <http://www.imss.gob.mx/imss-prospera>



- **Randomisation par appariement** : les individus ou les grappes sont regroupé(e)s par paires ayant des caractéristiques observables communes. Une unité de chaque paire est affectée aléatoirement au groupe expérimental tandis que l'autre unité est affectée au groupe contrôle. Cet appariement initial permet de garantir un certain équilibre et de réduire la taille d'échantillon requise.
- **Randomisation stratifiée** : lorsque des variables clés sont susceptibles d'influencer les résultats (p. ex., le niveau de revenus ou d'éducation), les participants sont partagés en groupes ou strates (p. ex., revenus faibles, intermédiaires ou élevés), et la randomisation est appliquée à chacune de ces strates. Cette méthode permet de répartir les principales variables de façon équilibrée entre le groupe expérimental et le groupe contrôle.

Une évaluation doit être menée pour suivre le processus de randomisation et vérifier qu'il est bien adopté (voir ci-après).

### 4. Collecter les données de référence sur les deux groupes

Avant ou après la randomisation des participants, les évaluateurs réalisent généralement un état des lieux pour collecter les données qui seront utilisées lors des comparaisons finales (voire lors de comparaisons ultérieures). Ces données de référence permettent aussi d'évaluer l'équivalence des caractéristiques initiales du groupe expérimental et du groupe contrôle. L'évaluation de l'équivalence permet de vérifier si la moyenne du groupe expérimental et la moyenne du groupe contrôle possèdent des variables observables en commun. Elle vise principalement à confirmer l'efficacité de la randomisation. Si des différences importantes sont trouvées (ou anticipées), le recours à la randomisation stratifiée peut être justifié.

Comme évoqué plus haut, la taille de l'échantillon étudié est estimée au moyen de calculs de puissance statistique. Généralement, l'état des lieux est réalisé au niveau des individus ou des ménages. Sont alors collectées des données relatives aux caractéristiques, à la situation socioéconomique, à l'éducation et à la santé des ménages, ainsi qu'à toute autre caractéristique potentiellement liée au programme devant être évalué et aux impacts qu'il vise à produire.

### 5. Collecter les données sur la mise en œuvre (voire sur les effets directs intermédiaires)

Les données relatives à la mise en œuvre doivent être collectées, idéalement au moyen d'une enquête intermédiaire qui se concentrera généralement sur le processus de mise en œuvre. Si elle ne survient pas trop tôt, cette étude peut aussi permettre les premières estimations sur l'impact du programme.

Il est important de vérifier que les membres du groupe contrôle ne font pas l'objet d'une « [contamination](#) », soit par la mise en œuvre d'une intervention similaire dans les zones couvertes par l'ECR, soit par un effet d'autocontamination (participants de l'étude qui passent d'un groupe à l'autre), ce qui contamine le processus de randomisation initial. Il faut aussi prêter attention au phénomène d'[attrition](#) entre les groupes qui risque de fausser les résultats. Par exemple, si les membres auprès desquels sont collectées les données relatives aux effets directs sont moins nombreux au sein du groupe contrôle que du groupe expérimental, les résultats risquent d'être biaisés (car les participants qui abandonnent le groupe ne sont pas pris en compte dans l'analyse).

### 6. Collecter les données sur les impacts

Une [étude finale](#) est réalisée après la mise en œuvre du programme. La décision relative à la date de fin du programme dépend de ce qui est prévu dans la théorie du changement concernant la durée nécessaire à l'obtention des impacts attendus. Lorsqu'un programme s'étend sur une longue période, les données finales peuvent être collectées après une période raisonnable, au-delà de laquelle on peut s'attendre à ce que les changements observés au niveau des effets directs de l'intervention commencent à être visibles.

Par exemple, dans le cadre d'un programme de nutrition incluant un enrichissement en fer, les chercheurs doivent veiller à respecter un délai suffisant pour permettre le déploiement du programme, une durée d'exposition suffisante pour garantir l'absorption du fer par l'organisme et l'amorce d'une baisse des cas d'anémie (liée à l'augmentation de la consommation de fer). Les données finales permettent d'estimer l'impact.

Lorsque la randomisation n'est pas optimale (comme mis en évidence par l'évaluation de l'équivalence évoquée plus haut), la méthode de la [double différence](#) (DID) peut être utilisée. Comme indiqué dans la Note n° 10, Présentation des méthodes de collecte et d'analyse de données dans l'évaluation d'impact, la méthode DID compare les changements observés au niveau des effets directs entre le groupe expérimental et le groupe contrôle. Elle permet aussi d'évaluer les effets hétérogènes et différentiels observés au sein des sous-groupes.

L'effet du programme peut dépendre de la composition du groupe : hommes ou femmes, riches ou pauvres, personnes instruites ou non, etc. La comparaison des effets directs observés chez ces différentes catégories, telles que représentées dans le groupe contrôle et les groupes expérimentaux, peut permettre d'estimer l'impact de l'intervention sur ces sous-groupes. Cette estimation peut être réalisée en appliquant la méthode de la double différence dans les différents sous-groupes de la population admissible, par exemple les hommes, les femmes et les enfants.

Selon la nature du programme, les évaluateurs peuvent aussi juger pertinent de réaliser une nouvelle étude après l'étude finale, afin d'estimer les effets à long terme de l'intervention. Le calendrier des études de suivi dépend de la rapidité à laquelle l'intervention peut générer des résultats. Par exemple, les transferts de denrées alimentaires peuvent produire des résultats relativement rapides sur le plan nutritionnel, tandis que les interventions visant à modifier les attitudes, les normes et les comportements nécessitent généralement plus de temps.

## 4. QUESTIONS ÉTHIQUES ET LIMITES PRATIQUES

Les ECR se heurtent à plusieurs difficultés d'ordre éthique et pratique, similaires à celles rencontrées dans le cadre des autres types d'évaluation (voir la Note n° 1, Présentation de l'évaluation d'impact). La présente section se limite toutefois aux questions spécifiques aux ECR.

Les questions éthiques spécifiques aux ECR sont liées à leur nature expérimentale qui nécessite de consulter les participants, afin d'identifier et de prendre en compte leurs souhaits, mais aussi de trouver un équilibre entre les risques et les avantages liés à l'essai. Les questions éthiques relatives à l'expérimentation sont d'autant plus importantes lorsque les ECR impliquent un groupe contrôle ne bénéficiant pas de l'intervention. En raison des inégalités qui découlent naturellement de ce type d'évaluation, il est très important que la randomisation soit transparente, particulièrement lorsqu'elle a lieu au niveau individuel. L'évaluateur doit s'assurer qu'aucune tension n'existe entre le groupe expérimental et le groupe contrôle.

Pour éviter ce risque, il est notamment recommandé d'expliquer clairement le but de la randomisation. La méthode de randomisation échelonnée peut aussi apaiser les préoccupations éthiques s'il est possible de garantir que l'intervention sera appliquée ultérieurement au groupe contrôle à condition d'être jugée efficace.

Le besoin initial qui pousse à réaliser un ECR soulève une autre question éthique. En l'absence de doute raisonnable sur les avantages et la rentabilité d'un programme, il n'est pas nécessaire de mener une évaluation approfondie (quelle qu'elle soit) ; un simple suivi de l'impact peut suffire à déterminer si le programme enregistre toujours les résultats attendus au fil du temps. Lorsque l'efficacité d'un programme n'est pas clairement établie, et que les ressources disponibles pour sa mise en œuvre sont limitées, les considérations éthiques peuvent toutefois justifier la randomisation des participants au programme, sans

perdre de vue le déploiement du programme auprès de l'ensemble de la population une fois son efficacité établie.

Il est aussi important de faire preuve de sensibilité lorsqu'il s'agit de collecter les données auprès du groupe contrôle. Les évaluateurs doivent veiller à solliciter les non-bénéficiaires de façon appropriée. Il est raisonnable de dédommager les participants à une enquête pour le temps qu'ils ont investi, tout en s'assurant que cette indemnisation n'affecte en rien les résultats (p. ex., en encourageant les individus à répondre d'une façon plutôt que d'une autre).

## 5. QUELLES SONT LES AUTRES MÉTHODES QUI FONCTIONNENT BIEN AVEC CELLE-CI ?

L'ECR est un modèle d'étude qui vise explicitement à répondre aux questions relatives à la causalité et à l'attribution. Son élaboration doit s'inscrire dans le cadre d'un plan d'évaluation plus général incluant aussi des méthodes de collecte de données (p. ex., observations, mesures ou entretiens directs) et d'analyse.

Avant de réaliser un ECR, il est recommandé d'effectuer une étude ou une évaluation formative de façon à évaluer si la mise en œuvre du programme peut être réalisée, voire améliorée.

Il est aussi souhaitable que l'ECR prévoie un outil permettant le suivi de la mise en œuvre, ou une évaluation de processus complémentaire. À la différence de l'ECR, qui s'intéresse à la situation contrefactuelle, l'évaluation de processus examine comment le programme a été mis en œuvre, généralement à partir de données quantitatives et qualitatives. L'évaluation de processus porte donc sur la chaîne de causalité et permet d'expliquer l'impact du programme. Les données collectées viennent compléter, voire confirmer ou infirmer, celles recueillies par le système de suivi spécifique au programme.

## 6. PRÉSENTATION DES RÉSULTATS ET ANALYSE

Afin de garantir une étude de qualité, l'ECR doit comprendre un mécanisme d'assurance qualité, et la préparation de la méthode et des constatations doit s'appuyer sur des données suffisamment détaillées. La description de l'ECR doit inclure la méthode employée et l'intervention visée par l'évaluation. Ces informations peuvent être tirées de l'évaluation de processus. La description détaillée de l'intervention permet de faire le lien entre la théorie du changement et l'analyse des constatations. Une description détaillée de la théorie du changement sera également fournie dans le cadre de la communication des constatations.

La description de la méthode employée doit aussi porter sur l'échantillonnage et la technique de randomisation, en précisant le nombre de grappes, de ménages et/ou d'individus qui composent le groupe expérimental et le groupe contrôle. En outre, le rapport doit inclure les tableaux de l'évaluation de l'équivalence (voir plus haut).

Les estimations sur l'impact peuvent être communiquées dans le cadre de l'analyse de la double différence, qui peut porter sur l'ensemble de l'échantillon ou sur les sous-groupes si l'on souhaite analyser les effets hétérogènes.

Le lien avec la théorie du changement doit ensuite être établi. L'analyse soutient-elle la théorie du changement ? Dans le cas contraire, quelle hypothèse sous-tendant la théorie du changement ne s'est pas concrétisée ? De même, les autres constatations tirées de cette étude doivent être expliquées. Quels facteurs internes et externes à la théorie du changement sont susceptibles d'avoir entraîné les résultats observés ? Cette analyse peut permettre aux évaluateurs de formuler des constatations concrètes et pertinentes sur le plan politique, qui seront présentées dans la conclusion du rapport en lien direct avec l'analyse de données. Généralement, il est aussi important d'inclure une réflexion sur la possibilité d'extrapoler les résultats dans différents environnements, en précisant lesquels le cas échéant.

## 7. EXEMPLE DE BONNES PRATIQUES

Un projet pilote d'intervention en faveur de la petite enfance financé par l'UNICEF<sup>4</sup> a été mené au Pakistan en vue d'évaluer l'efficacité et la faisabilité de la mise en œuvre d'interventions axées sur la petite enfance visant à renforcer et à améliorer la santé des jeunes enfants. Afin de mettre en œuvre les différents volets de l'intervention, ce projet pilote s'est appuyé sur le programme des travailleuses de la santé (Lady Health Worker - LHW) déjà en place. Randomisée en grappes et dotée d'un plan factoriel, l'étude s'est déroulée dans les zones rurales du Pakistan.

Plusieurs grappes (définies d'après les zones couvertes par LHW) ont été réparties aléatoirement entre différents groupes : le **groupe contrôle**, qui a reçu les services essentiels de nutrition et de santé fournis par le programme LHW ; le **groupe « amélioration de la nutrition »**, qui a reçu des conseils en matière de nutrition et d'alimentation active, et des suppléments nutritionnels (Sprinkles®) destinés aux jeunes enfants âgés de 6 à 24 mois ; le **groupe « développement de la petite enfance »**, qui a reçu des conseils et des soins visant à favoriser et à stimuler le développement de l'enfant, notamment par le biais d'un accompagnement, de visites mensuelles à domicile et de réunions de groupe ; et le **troisième groupe expérimental**, qui a bénéficié d'un ensemble d'interventions axées à la fois sur le développement de la petite enfance et l'amélioration de la nutrition. Les interventions ont été menées auprès de chaque famille ayant un jeune enfant de moins de 24 mois, située dans la zone desservie par LHW.

La santé des enfants a été évaluée à partir d'un ensemble de données et d'effets directs enregistrés à plusieurs reprises auprès des enfants et de leurs familles jusqu'aux 24 mois de l'enfant. Afin de réduire le risque de partialité au moment de l'évaluation, l'équipe chargée de la collecte des données n'avait aucun contact avec l'équipe chargée du soutien ni aucune information concernant le type d'intervention dont l'enfant avait bénéficié. L'étude a montré qu'à l'âge de 12 mois, les enfants des trois groupes expérimentaux présentaient des capacités cognitives, langagières, motrices et socioémotionnelles nettement plus avancées que ceux du groupe contrôle. Les enfants du groupe combiné (ayant bénéficié d'interventions en faveur du développement de la petite enfance et de l'amélioration de la nutrition) présentaient des résultats cognitifs et langagiers bien meilleurs que ceux du groupe uniquement axé sur l'amélioration de la nutrition. À l'âge de 24 mois, les enfants des trois groupes expérimentaux présentaient tous des capacités cognitives, langagières et motrices plus élevées que ceux du groupe contrôle, à plus forte raison dans les deux groupes ayant bénéficié d'interventions axées sur le développement de la petite enfance.

Ce projet pilote constitue un bon exemple d'utilisation de l'ECR, tant au niveau de la prise en compte de la situation que de la méthode de mise en œuvre et d'utilisation.

Le programme était bien défini et assorti d'objectifs clairs. Les communautés admissibles ont été identifiées et réparties aléatoirement entre les groupes expérimentaux et le groupe contrôle. La collecte et l'analyse des données reposant sur une solide théorie du changement, les évaluateurs sont parvenus à comprendre comment les objectifs de l'intervention ont pu être atteints.

## 8. EXEMPLES DE DIFFICULTÉS

**Préserver l'intégrité de l'évaluation** : malgré le recours à la randomisation, certaines difficultés sont parfois inévitables, notamment les suivantes : (1) faible taux d'utilisation de l'intervention ; (2) non-respect des procédures prévues ; (3) autocontamination ou contamination du groupe contrôle par d'autres interventions visant les mêmes effets ; et (4) modification du programme évalué ou de sa couverture

<sup>4</sup> Yousafzai, A. K. *et al.*, 2012, « The Pakistan Early Child Development Scale Up (PEDS) Trial: Outcomes on child development, growth and health », rapport sur les données relatives aux effets directs du projet pilote de l'intervention en faveur de la petite enfance menée au Pakistan, UNICEF Pakistan

géographique. La plupart de ces problèmes peuvent être évités lors de la phase d'analyse, à condition que les évaluateurs collectent les données nécessaires pour identifier ces problèmes et les résoudre.

Le faible taux d'utilisation d'une intervention désigne une situation dans laquelle les programmes n'intéressent pas les bénéficiaires visés, ou ne sont pas bien compris par ceux-ci. Par exemple, les programmes d'assurance se caractérisent souvent par une faible utilisation. Des données qualitatives sont souvent nécessaires pour expliquer cette faiblesse.

À titre d'exemple sur le non-respect des procédures prévues, un ECR mené en Chine<sup>5</sup> a montré qu'un programme visant à fournir des lunettes à des étudiants de l'enseignement secondaire avait aussi entraîné une augmentation de l'utilisation des lunettes chez certains membres du groupe contrôle. Une enquête a permis d'établir que les médecins chargés des examens ophtalmologiques avaient donné des paires de lunettes non utilisées à des étudiants du groupe contrôle. Cet exemple est un cas d'autocontamination. Le recours à la méthode de l'appariement a ensuite permis d'exclure les paires du groupe contrôle ayant été contaminé.

**Méthode en grappes - absence d'ajustement fondé sur les erreurs types :** cette erreur technique courante augmente artificiellement la puissance statistique de l'évaluation et peut entraîner la formulation de conclusions erronées sur l'efficacité du programme. Prenons le cas d'une étude pour laquelle plusieurs écoles ont été réparties aléatoirement entre différents bras expérimentaux. Si la phase d'analyse (au cours de laquelle les effets directs en termes d'apprentissage sont comparés au moyen de tests menés chez les enfants des groupes expérimentaux et du groupe contrôle) n'inclut pas le contrôle de la composition des grappes, l'impact risque d'être surévalué. Tous les logiciels statistiques utilisés pour réaliser des évaluations d'impact prévoient cet ajustement, qui doit être systématiquement effectué par les chercheurs.

**Importance excessive accordée à l'effet moyen de l'intervention :** l'ECR estime objectivement l'effet moyen d'un programme. Toutefois, les décideurs politiques s'intéressent rarement à cette moyenne et préfèrent recevoir des informations relatives à l'efficacité du programme pour des sous-groupes particuliers, notamment dans le cas des programmes axés sur les questions d'équité. Par exemple, l'évaluation d'un programme américain d'intervention en faveur de la petite enfance, intitulé Early Head Start<sup>6</sup>, a montré que le programme, bien qu'efficace « en moyenne », avait en réalité eu des effets néfastes sur les familles les plus vulnérables. Pour cette raison, la préparation de rapports axés sur cet effet moyen risquait d'induire en erreur les décideurs politiques et les prestataires de services. Pour pallier ce problème, il est possible de détailler les résultats des différents sous-groupes sur lesquels l'intervention est susceptible d'avoir un impact différent. Toutefois, ce type de méthode doit être identifié dès la phase d'élaboration, et les calculs de puissance statistique adaptés en conséquence.

**Rejet de la randomisation :** il est fréquent que le personnel de l'organisme d'exécution conteste le recours à la randomisation. L'engagement des cadres supérieurs ne suffit pas toujours à garantir celui du personnel de terrain. C'est ce que montre l'étude du programme de fourniture de lunettes mis en œuvre en Chine, au cours duquel les médecins ont donné des paires de lunettes à des membres du groupe contrôle.

---

<sup>5</sup> Glewwe, P. *et al.*, 2006, « The Impact of Eyeglasses on the Academic Performance of Primary School Students: Evidence from a Randomized Trial in Rural China », rapport de conférence, Université du Minnesota et Université du Michigan. Disponible à l'adresse : <http://ageconsearch.umn.edu/bitstream/6644/2/cp06ql01.pdf>

<sup>6</sup> Westthorp, G., 2012, « Using complexity-consistent theory for evaluating complex systems », *Evaluation*, 18 (4), p. 405-420

## 9. PRINCIPALES LECTURES ET LIENS UTILES

Ambroz, A. et Shotland, M., 2013, « Randomized Control Trial (RCT) », page Internet, BetterEvaluation. <http://betterevaluation.org/plan/approach/rct>

Bloom, H., 2006, « The Core Analytics of Randomized Experiments for Social Research », MDRC Working Papers on Research Methodology, MDRC, New York. Disponible à l'adresse suivante : [http://www.mdrc.org/sites/default/files/full\\_533.pdf](http://www.mdrc.org/sites/default/files/full_533.pdf)

Duflo, E. *et al.*, 2006, « Using Randomization in Development Economics Research: A Toolkit », département d'Économie du Massachusetts Institute of Technology et centre de recherche Abdul Latif Jameel Poverty Action Lab, Cambridge. Disponible à l'adresse : <http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf>

Gertler, P. J. *et al.*, 2010, L'évaluation d'impact en pratique in Practice, Banque mondiale, Washington D.C. Disponible à l'adresse : <http://issuu.com/world.bank.publications/docs/9780821387528/16>

Glennerster, R. et Takavarasha, K., 2013, *Running Randomized Evaluations: A Practical Guide*, Princeton University Press, Princeton

Raudenbush, S. W. *et al.*, 2011, Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01), logiciel. Disponible à l'adresse suivante : [www.wtgrantfoundation.org](http://www.wtgrantfoundation.org)

Spybrook, J. *et al.*, 2011, Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software Version 3.0, documentation logicielle. Disponible à l'adresse suivante : [www.wtgrantfoundation.org](http://www.wtgrantfoundation.org)

Urbaniak, G. C. et Plous, S., 2013, Research Randomizer (Version 4.0), logiciel de randomisation et d'échantillonnage aléatoire. Disponible à l'adresse suivante : <http://www.randomizer.org/>

White, H., 2013, « An introduction to the use of randomised control trials to evaluate development interventions », *Journal of Development Effectiveness*, 5 (1), p. 30-49. Disponible à l'adresse suivante : <http://www.tandfonline.com/doi/pdf/10.1080/19439342.2013.764652>

## GLOSSAIRE

<b><u>Analyse de dépendance</u></b>	<i>Analyse basée sur un enchaînement de causalité (aussi appelé cadre analytique) qui décrit les relations directes et indirectes entre les variables indépendantes (interventions) et dépendantes (produits/effets directs/impacts).</i>
<b><u>Appariement sur le score de propension</u></b>	<i>Méthode quasi expérimentale qui associe des individus/ménages du groupe expérimental avec des individus/ménages comparables du groupe témoin, pour ensuite calculer la différence moyenne dans les indicateurs pertinents.</i>
<b><u>Attrition</u></b>	<i>Situation dans laquelle les participants d'une étude quittent l'échantillon entre deux cycles de collecte de données. Le phénomène d'attrition peut menacer la validité interne d'une étude en modifiant la composition de l'échantillon.</i>
<b><u>Contamination</u></b>	<i>Inclusion d'un ou plusieurs sujets dans un groupe expérimental, alors qu'ils ne sont pas représentatifs de la population ou ne sont pas censés bénéficier du traitement/de l'intervention. La contamination peut se produire lorsque les participants/sujets d'un groupe contrôle bénéficient d'un traitement/d'une intervention sans que cela ne soit prévu. Ce phénomène entraîne une réduction des effets du traitement/de l'intervention décrits dans les résultats de la recherche.</i>
<b><u>Double différence (DID)</u></b>	<i>Aussi appelée la méthode de la « différence de différence ». Méthode qui évalue l'impact en comparant les changements en matière d'effets directs observés au fil du temps entre le groupe expérimental et le groupe témoin.</i>
<b><u>Étude finale</u></b>	<i>Étude réalisée à la fin de l'intervention, généralement en vue de comparer les résultats à l'étude de référence. Terme associé : étude de référence</i>
<b><u>Groupe expérimental</u></b>	<i>Sujets/participants exposés à la variable indépendante.</i>
<b><u>Méthode d'encouragement</u></b>	<i>Modèle expérimental dans le cadre duquel la randomisation des participants ne porte pas sur une intervention, mais sur l'accès à des documents promotionnels ou à un supplément d'information sur les avantages de l'intervention à leur disposition, en vue de renforcer la participation à cette intervention. Les méthodes d'encouragement peuvent être utilisées lorsque des programmes sont accessibles à tous, mais qu'ils ne sont pas adoptés par tous. Voir : méthode expérimentale</i>
<b><u>Méthode expérimentale (ECR)</u></b>	<i>Modèle de recherche ou d'évaluation constitué d'au moins deux groupes sélectionnés de façon aléatoire (un groupe expérimental et un groupe contrôle) dans lesquels le chercheur teste ou introduit une intervention (un nouveau programme ou une nouvelle politique par exemple) et mesure son impact sur la variable dépendante au minimum à deux reprises (mesures avant et après le test). Issus d'environnements cliniques et connus comme étant la « norme d'excellence » de la recherche médicale et de la santé, les ECR sont souvent utilisés, notamment pour traiter les questions de recherche évaluative visant à mesurer l'efficacité d'interventions programmatiques et politiques dans des contextes de développement.</i>

<b><u>Méthode non expérimentale</u></b>	<i>Méthode de recherche qui n'inclut pas de groupe contrôle ou de groupe témoin et/ou qui ne s'appuie pas sur une évaluation initiale. Ainsi, plusieurs facteurs empêchent l'attribution d'un effet observé à l'intervention. Voir : méthode expérimentale, méthode quasi expérimentale.</i>
<b><u>Méthode quasi expérimentale</u></b>	<i>Méthode de recherche/d'évaluation qui n'alloue pas de façon aléatoire certaines conditions de traitement aux participants, mais dont les groupes témoins sont constitués à l'aide de moyens statistiques. Elle diffère de l'expérimentation (classique) contrôlée, car il n'y a pas de distribution aléatoire du traitement/de l'intervention.</i>
<b><u>Modèle d'ECR en grappes</u></b>	<i>Modèle expérimental dans lequel chaque unité d'étude est composée de plusieurs unités bénéficiant de l'intervention, plutôt que d'un seul sujet. Par exemple : les interventions éducatives sont généralement mises en œuvre dans des écoles entières, même si l'intervention est en réalité axée sur l'enseignant, la classe ou chaque élève et que les effets sont mesurés à l'échelle de ce dernier. Voir : essai contrôlé randomisé (ECR)</i>
<b><u>Reconstitution de processus</u></b>	<i>Étude de cas visant à identifier l'« inférence causale » en élaborant des hypothèses alternatives, puis en recueillant des données probantes (indices) pour déterminer si elles sont ou non compatibles avec les hypothèses disponibles.</i>
<b><u>Théorie du changement</u></b>	<i>Explique comment les activités s'entendent comme produisant un ensemble de résultats qui contribuent à la réalisation des impacts finaux prévus. Elle peut être élaborée pour tous les niveaux d'intervention : événement, projet, programme, politique, stratégie ou organisation.</i>